Modelling, Self and Consciousness: Further Perspectives of AI Research

Ricardo Sanz
Universidad Politécnica de Madrid, Spain
Alexander Meystel
Drexel University, USA

Abstract:

Sound measurement of intelligence cannot be reduced to just measurement of performance. It is necessary to measure the real capabilities of the behavior generation engine of machines to be able to determine with precision their suitability for any particular task. We will see that it is necessary to focus on the architecture of the systems. The paper will present a summarial description of inner machinery of intelligence and how this architecture can serve as the basement for higher mental functionality. This will lead us to the formulation of a theory of conscious behavior and a the proposal of a research program focused into the nature and mechanisms of machine consciousness.

Keywords:

Intelligence, performance, mental architecture, self, consciousness.

1 Introduction: Measuring What ?

As Lord Kelvin said, *"to measure is to know´´* and hence the importance of measuring intelligence to know better about it. Obviously, we know that the problem is not easy just from the very beginning: Is "Measuring of Intelligence´´possible ?

In our search for machine replacements of humans in boring, dangerous or economically unviable activities, we use to check performance of machines [$Mind_M$] against performance of humans [$Mind_H$] or cognitive models of humans [$Mind_C$]. However, do we sufficiently understand ourselves [$Mind_H$]? or do we sufficiently understand the systems we design, manufacture, and use [$Mind_M$]?

Generally speaking, we can consider two ways of measuring intelligence: with regard to a particular task or independently of any particular task (Pease [1] refers to this last form as *a priori* intelligence).

Intelligence manifests itself in the autonomous successful performance of tasks. Or, to be more precise, in the autonomous successful performance of a *task* by a specific *agent* in a concrete *context* [2].

```
Autonomy[AGENT,TASK,CONTEXT]
```

Extending the base idea of measuring *a priori* intelligence we need to measure this faculty independently of the concrete task, the concrete context and the concrete agent; otherwise what we will have is a concrete, particular measure, not very helpful to compare systems with a wide application domain (this being the case of conventional IQ tests, that just measure the capability of performing these tests and where extrapolation of results to other activities is highly risky).

To be able to obtain a measure of pure (*a priori*) intelligence we need to eliminate from the equation such factors as concrete bodies, concrete tasks and concrete contexts. This will leave the pure essence of intelligence. This vision of intelligence matches our intuitive, abstract notion of intelligence as a central faculty independent of particular factors that surround the activity. How can this possibly be achieved ?

In this paper we will try to identify the core essence of intelligence to be able to *directly* measure its capabilities instead of measuring the result of these capabilities in a concrete task. The conclusion of this identification will lead us to a research program that re-gains that old dream of artificial intelligence: building conscious machines.

2 Architecture and Performance

Presuming a functional equivalence of basic building materials, all our theories of mind [$Mind_H$], [$Mind_M$] and [$Mind_C$] lead us to the concussion that only mental architecture can account for intelligent systems performance.

The perceived intelligence is strongly correlated with success. Intelligent systems architecture is a critical factor for success in the performance of any task [3]. Architecture is hence the point to focus our search for an a priori measure of

intelligence [4]. Bad architectures lead to non performing systems.

Ignoring sterile differences between reactive and deliberative intelligence, all these theories do constitute interpretations that depend critically on representation of goals, states, contexts and bodies [5].

That representation is a central factor of intelligent performance has been know for decades. Execution engines do exploit representations to derive agent's actions. These execution engines receive varying names depending on the concrete task at hand: planners, behavior generators, predictors, etc. All them exploit the information about the world stored in a model (a world model) to derive actions. Minds are control systems based on models.

This leads to a typical architectural pattern for representation-based control system: the elementary loop of functioning *perceive-represent-plan-act* that is used as an elementary building block for more complex architectures (see Figure 1).

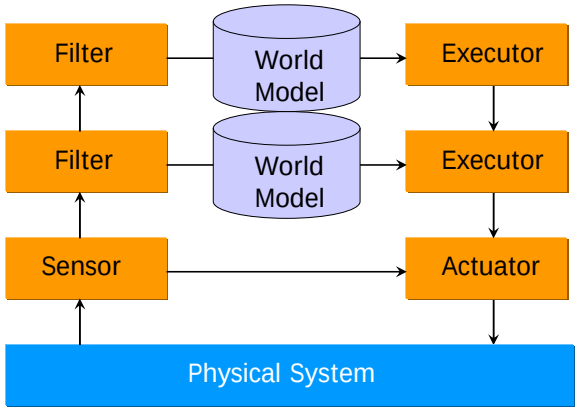

Figure 1: A basic, two layered, model-based control architecture. Each layer constitutes an elementary loop of functioning.

The effectivity of a concrete pair of components [*model*,*engine*] depends on the particular factors for autonomy mentioned before: *task*, *context* and *agent*. This means that a concrete pair, for example *[ordinary differential equations, Runge-Kutta simulator]* can be better than other pair *[first order logic predicates, resolution engine]* for a concrete task, for example *tank temperature prediction*, in a particular context, for example a *well-engineered refinery*, for a specific agent, for example a *model-based predictive controller*.

In many cases, this specificity lead us to sacrifice generality when dealing with constraints to attain specific execution properties (speed, robustness, cost). For example, mutiresolutional representation and control hierarchies offer cost effective solutions with bounded resources; for speed enhancement, compiled representations and engines adapted for them are employed.

Generality, however, is an extremely desirable property for a pair *[representation, engine]*. Generality is the mark of pure intelligence. Tradeoffs do obviously exist and have been used -mistakenly- as arguments against the suitability of general representations for the construction of intelligent agents [6].

Generality is out of questioning, however, because if we want to give to our systems control mechanisms with a high degree of *a priori* intelligence we need generality to overcome the barriers of the three factors: task, context, agent.

The broader the set of solvable tasks the greater the intelligence of the machine. This was that old dream of the Ultimate Problem Solver. For example: a washing machine is more intelligent if it is also able to minimize water consumption.

The greater the context-independence of the controller the higher the intelligence of the machine. This means that the controller can reach its objectives in a variety of execution contexts, *i.e.* is robust against variations in its execution environment, being able to handle uncertainty in a proper way. For example, a transelevator in an automated warehouse is more intelligent if it can avoid people eventually obstructing its way.

And last, machines that tolerate alterations in their own bodies and still fulfil their objective are more intelligent. For example, a car controller that can maintain car stability with a broken tire is more intelligent.

3 Systems with Self

Evolutionary pressure has forced a race-of-brains in the biosphere. One of the highest advances is when systems become able to extend its own capabilities. This means that the system can go beyond current engines and representations, current contexts and tasks, and even its current body. The highest levels of intelligence are those that not only do learning (enhancing engines) but also modelling (enhancing representations).

Intelligent natural systems do learn autonomously, or are provided externally the units of knowledge that are required for their successful functioning. This last approach is simpler in the case of artificial systems with limited scope. Anticipatory systems [7] can do all this process autonomously. The results of knowledge acquisition are models that capture reality (with the precision and level of resolution needed for the task). These models can be —easily— organized into automata models and hence the usefulness of computers to implement intelligent systems.

Truly intelligent systems do have models of their world and the tasks they perform and are able to enhance them. In fact, this is what all the business of science is about. Better models of reality to overcome, with our technologies, all the barriers

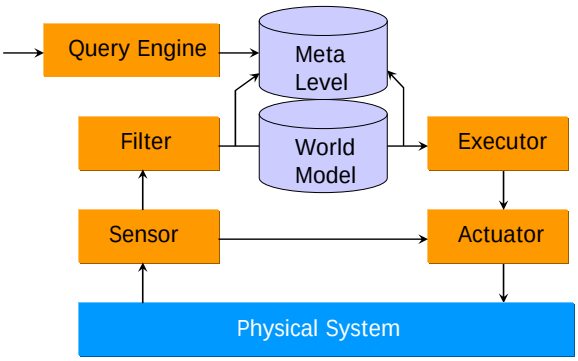

Figure 2: A not so basic, two layered, model-based control architecture. The metalevel provides introspection over an elementary loop of functioning.

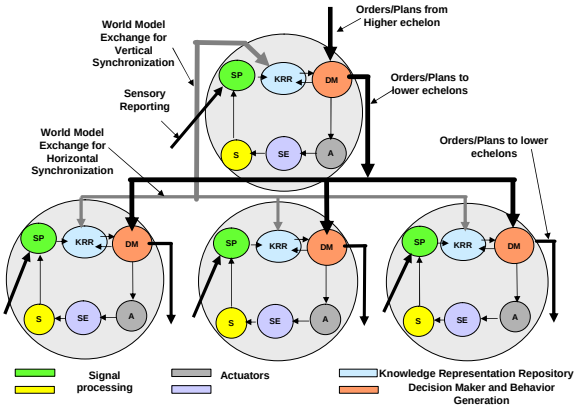

Figure 3: Another not basic, layered, model-based control architecture. The multiresolutional heterarchy goes beyond the single elementary loop of functioning into a collection of interacting loops.

from the past. Based on our better models of the external world (the context) we can do things (tasks) that go far away of what our grandfathers were able to do (with a similar body)[1].

The last step is easy to see: Truly intelligent systems do also model themselves. True intelligent systems maintain continuously updated, continuously enhancing representations of themselves. True intelligent systems are self-conscious.

Using this internal representations of themselves, intelligent machines can use their reasoning engines to reason about themselves and act accordingly. This representation and reasoning do also include more basic representation and reasoning processes (see Figure 2); intelligent systems do have metalevel representations and reasoning systems that, coupled with a query engine and a language interface, are used to interchange mental states with others: the states of the representations and reasoning processes (see Figure 3).

This simple analysis capture a commonsense thought: To be truly effective, intelligent systems need to be aware of themselves, i.e. need to be self-conscious. The nature of the self is this continuous perception and control of the body of the agent (see Figure 4). This is *the ghost in the machine*.

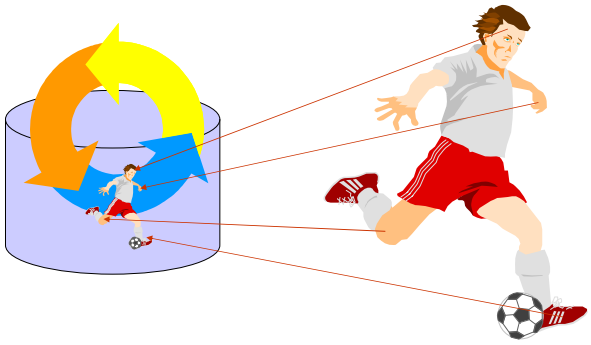

Figure 4: The nature of the self: the model of the body inside the model of the world and a loop of model-based control over it.

4 Autonomous Performance of Systems with Self

Automata models used by intelligent systems consist of explanations of the environment, states of the system and the appropriate rules of action. The basic process implemented in the loop of functioning is reproduced in Figure 5.

This model of reality need not be unitary but composed by a collection of elementary models. The set of all these models is aggregated together by the representation system of the agent.

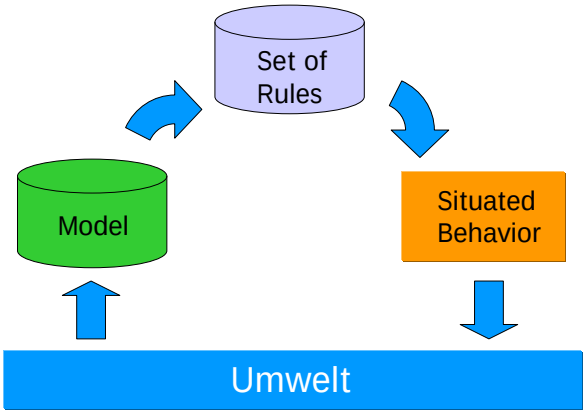

Figure 5: The process of modelling the world to behave inside it generates a semiotic loop.

[1] Consider, for example, the increase in life expectancy in the last fifty years based on better models of human inner workings.

Representation as a system of models appears in both natural and constructed systems.

This representation system also includes intelligent system goals that are distributed across the heterarchy of models of reality. In a similar way, we can think that autonomy is distributed over the architecture with the same level of granularity. Each level has a component of autonomous behavior determined by one or many goals of this level and the corresponding performance measures, and as the resolution becomes lower, both behaviors and goals are generalized. Autonomy is embodied in the hierarchy of goals and performance measures supported by internal models. In some sense, we can say that automata models have a rudimentary form of free will.

The intelligent system has a certain degree of autonomy. However, its activity is oriented toward the goal of the larger system (lower resolution unity that the intelligent system belongs to).

Consider a bee. While everybody would agree about its autonomy, no one will doubt that it is —the autonomous activity— oriented toward the goal of the swarm. At least, it must not contradict it in the long term. We also can talk about the autonomy of a can-picking robot. But one should agree that its autonomy is oriented toward the goal implanted into this robot by the designer —a sort of lower resolution level.

In the case of artificial systems, this dependency between levels of autonomy is so strict, that we strive only for *bounded autonomy* in the design of these systems.

Actually, the mechanisms of autonomy in the overall system as well as functioning of autonomic subsystems determine the viability of the system.

At each level of resolution, one can talk only about the degree of autonomy enclosed within the goal-oriented activity. This degree is determined by the self of this particular level. The degree of autonomy is required to cope with the high-resolution eventualities that at the lower resolution level are parts of the uncertainty taken into account.

The question *Why do we need systems with "self´´ ?* has a clear answer now: Having a self (a continuously updated model of agent's body) do increase intelligent systems performance.

To be more concrete, this increased performance can be shown by the collection of tasks that depend on this schema of representation and control:

- Introspection/reportability
  - Optimization (reorganizing own inner processes)
  - Socialization (collaborating with other agents)
- Fault handling
  - Fault detection and isolation (finding problems)
  - Diagnosis (identifying causes)
  - Fault management (devising walkarounds)
- Autonomous behavior
  - Regoaling (changing tasks)
  - Reconfiguration (changing body)
  - Tooling (changing context)

5 Intelligence, Self, Consciousness

In this process of intelligent systems analysis there are many objectives. In some sense we are stepping toward $[Mind_H]$ by means of implementations of $[Mind_M]$ that provide progressively accurate behavior. To our understanding, the architectural model presented here is a good unified theory of natural and artificial intelligence systems.

This theory do explain some of the more difficult observed aspects of human minds, while avoiding entering into the non-implementable field of metaphysics.

One of the more puzzling aspects for human mind is the uniqueness of "self". As we have seen, the model of the extended-plant (the body+the controller) can be single or multiple, not necessarily unitary. The question is: Is there any reason for the uniqueness of the "self"?.

Our hypothesis is that unitary selves provide evolutionary advantages (better autonomous performance). Having a single self enables resolution of autonomous control problems with scarce resources in the presence of higher degrees of uncertainty. The presence of the single self guarantees the coherence of the set of multiresolutional goals of the intelligent system.

In relation with $[Mind_H]$, there are no widely accepted explanations of conscious/unconscious behavior. Some authors distinguish between unitary and dual explanations (having one or two mechanisms for the conscious and the unconscious) and multiple theories are available on both sides (See [8] for a good survey).

One of the major problems is that while unitary explanations are more aestetically pleasant, they fail to provide the necessary qualitative distinction between the conscious and the unconscious that many authors want to see. The origin of the problem can be traced to the perceived distance between conscious and unconscious that from our point of view does not exist. Consciousness is not a binary property, it only looks like that because the interaction with external agents (with others) is performed only an a concrete high level of the control hierarchy.

The main obstacles for a unified theory of mind ($[Mind_H]$+$[Mind_M]$ are the chauvinism of human species and the manicheism of most theories, that appears everywhere: Representation/representationless, deliberative/reactive conscious/unconscious, symbolic/subsymbolic, biologic/beer can.

There are no intelligent systems without representation; every system that has a sensor has representation. Reactive control systems are just degenerate cases of deliberative control systems (where deliberation is reduced to a simple I/O mapping).

There is even disagreement about what is *consciousness*. Some authors distinguish three types: Access consciousness, reflective consciousness and phenomenal consciousness. There are authors that distinguish upon seven types !!!.

The key for the emergence of self-consciousness is integration of information about the body with information about the world (see Figure 6)[9].

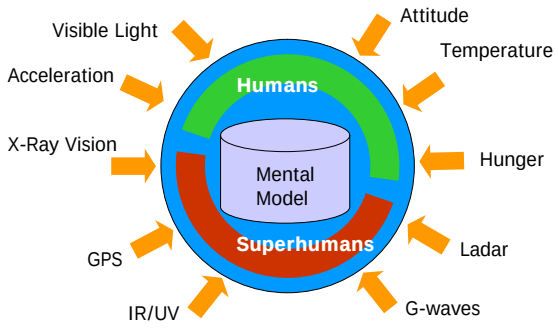

Figure 6: The process of integrating information from incoming sensors lead to progressive world-awareness. Systems with better sensors can be more aware if properly designed.

Consciousness increases as more information (from the outside, from the inside, from the mental processes, etc.) is integrated in a dynamical mental model that includes the self.

The -wicked- problem is not achieving consciousness but achieving human-like consciousness (i.e. being recognized as humans by other human minds) and this can be done only by means of human-like reportability and a human-like mental architecture. But building humans is nonsense from a practical perspective and recognizing consciousness in very alien systems is something that not everybody is prepared to do. As Thomas Nagel would ask: *What is it like to be a Tomahawk missile ?* [10]

6 Conclusions

Semiotic principles of computation provide a sufficient background for developing computer systems that demonstrate elements of self, consciousness, and free will.

The prerequisite for achieving this milestone in the intelligent computer systems development is focusing on interpretability of representations as the major factor of performance. This can be approached by developing multiresolutional (multigranular, multiscale) systems of knowledge representation equipped with a sufficient set of procedures to exploit the representations.

Eventually, the integration of these hierarchical model-based control loops will lead to ascertain the emergence of SELF.

We can conclude that *consciousness* is just an operational mode of a -not necessarily- complex controller. Being conscious is just having a running controller (being ON). Self-consciousness appears by the very same method when the sensed plant is the intelligent system proper.

As the multigranular system of semiotic closures emerges, it arrives at the phenomenon of representing and monitoring itself: self-consciousness. SELF [consciousness] is the multigranular system of semiotic closures constructed in representation for interpreting intentionalities of a system within its blended global multiscale coordinates. SELF is also supported by a multigranular system of goals that can be considered a provisional reminder of the results earlier produced by its intentionality system.

References

[1] A. Pease, “Evaluation of intelligent systems: The high performance knowledge bases and ieee standard upper ontology projects,” in *Proceedings of the Workshop Measuring the Performance and Intelligence of Systems, PerMIS’2001*, (Mexico D.F.), September 4 2001.

[2] R. Sanz, F. Matía, and S. Galán, “Fridges, elephants and the meaning of autonomy and intelligence,” in *IEEE International Symposium on Intelligent Control, ISIC’2000*, (Patras, Greece), 2000.

[3] J. Albus and A. Meystel, *Engineering of Mind: An Introduction to the Science of Intelligent Systems*. Wiley Series on Intelligent Systems, New York: Wiley, 2001.

[4] H. H. Hexmoor, “Smarts are in the architecture!,” in *AAAI Spring Symposium*, 1995.

[5] A. Meystel and J. Albus, *Intelligent Systems: Architecture, Design, Control*. Wiley Series on Inteligent Systems, New York: Wiley, 2001.

[6] R. A. Brooks, “Elephants don’t play chess,” *Robotics and Autonomous Systems*, vol. 6, pp. 3–15, 1990.

[7] R. Rosen, *Anticipatory Systems*. Pergamon Press, 1985.

[8] J. G. Taylor, *The Race for Consciousness*. Cambridge, MA: MIT Press, 1999.

[9] R. Sanz, “An integrated control model of consciousness,” in *Proceedings of the conference Toward a Science of Consciousness*, (Tucson (AZ), USA), April 8-12 2002.

[10] T. Nagel, “What is it like to be a bat?,” *The Philosophical Review*, October 1974.

[11] M. Asada, E. Uchibe, and K. Hosoda, “Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development,” *Artificial Intelligence*, vol. 110, no. 2, pp. 275–292, 1999.

[12] A. Sloman, “Virtual machines and consciousness,” technical report, University of Birmingham, 2002. http://www.cs.bham.ac.uk/research/cogaff/.

[13] C. Joslyn, “Levels of control and closure in complex semiotic systems,” in *Workshop on Closure: Emergent Organizations and their Dynamics*, (University of Ghent, Belgium), May 3-5 1999.

[14] N. R. Jennings, “On agent-based software engineering,” *Artificial Intelligence*, vol. 117, pp. 277–296, 2000.

[15] N. Shadbolt, “The shape of things to come,” *IEEE Intelligent Systems*, pp. 2–3, September/October 2001.

[16] A. Meystel and R. Sanz, “Self-identity of complex control systems,” in *Proceedings of the Workshop on Theoretical Fundamentals of Intelligent Systems*, (Durham (NC), USA), March 11 2002.

[17] W. Dress, “A Bayesian approach to extracting meaning from system behavior,” in *Proceedings of IEEE Systems, Man and Cybernetics Conference*, 1998.

[18] P. Horn, “Autonomic computing: Ibm perspective on the state of information technology.” IBM Research, 2001.